

## ПРИКЛАДНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 004.912

С.Ф. Липницкий, А.А. Мамчич, С.А. Сорудейкина

ВЕБ-ПОИСК И АННОТИРОВАНИЕ НАУЧНО-ТЕХНИЧЕСКОЙ  
ИНФОРМАЦИИ НА ОСНОВЕ ТЕМАТИЧЕСКИХ КОРПУСОВ ТЕКСТОВ

*Предлагаются алгоритмы поиска в Интернете текстовых документов и их аннотирования. Разрабатывается архитектура системы веб-поиска и аннотирования научно-технической информации. Рассматриваются состав и структура лингвистических словарей базы знаний, используемых при индексировании текстовых документов и синтезе аннотаций.*

**Введение**

Аннотация играет важную роль в системе сведений о текстовом документе: она дополняет его библиографическое описание и содержит ряд характеристик, которые дают первичное представление о публикации наряду с другим информационным жанром – рефератом. Различие между этими жанрами состоит в том, что «реферат включает краткое, максимально свернутое изложение содержания публикации, а аннотация – краткую ее характеристику» [1]. Реферат используется для предварительного ознакомления с текстовым документом без обращения к первоисточнику, а аннотация имеет пояснительный или рекомендательный характер [2].

В существующих программных системах индексирования, поиска и постпоисковой обработки текстовых документов применяются главным образом технологии, ориентированные на исследование структуры и статистических характеристик самих документов без привлечения дополнительной информации [3–6]. В настоящей статье эти задачи решаются с использованием тематических корпусов текстов и сформированных на их основе словарей базы знаний. Предложенные алгоритмы веб-поиска и аннотирования научно-технической информации отличаются универсальностью, т. е. независимы от предметной области. Настройка системы на конкретную предметную область сводится к созданию соответствующего тематического корпуса текстов и актуализации словарей базы знаний.

**1. Архитектура системы веб-поиска и аннотирования текстовых документов**

Функциональными компонентами проектируемой системы веб-поиска и аннотирования текстовых документов являются три подсистемы (рис. 1):

- автоматизированное рабочее место (АРМ) эксперта-лингвиста;
- подсистема поиска текстовых документов в Интернете;
- подсистема аннотирования найденных документов.

**1.1. Автоматизированное рабочее место эксперта-лингвиста**

АРМ эксперта-лингвиста – это программно-информационный инструментарий, предназначенный для визуализации процедур создания и ведения баз данных и знаний и управления этими процедурами.

В базе данных эксперт-лингвист накапливает и классифицирует по тематическим разделам предметной области электронные варианты текстовых документов для последующего создания на их основе тематических корпусов текстов. (Тематический корпус текстов – это совокупность полнотекстовых документов, посвященных какой-либо конкретной тематике.) В базу данных помещаются также электронные версии двуязычных словарей (для всех языков, документы на которых обрабатываются системой), электронные толковые словари и энциклопедии, касающиеся предметной области. В ней могут быть также представлены варианты готовых ан-

нотаций. На основе информации из базы данных в системе формируются лингвистические словари базы знаний.



Рис. 1. Архитектура системы веб-поиска и аннотирования текстовых документов

### 1.2. Подсистема веб-поиска текстовых документов

Основными функциями подсистемы веб-поиска являются индексирование веб-страниц на основе использования словарей базы знаний и поиск текстовых документов по запросам пользователей. Основными компонентами подсистемы являются:

- программы загрузки веб-страниц «Spider» («паук») и «Crawler» («путешествующий паук»). Это браузероподобные программы, предназначенные для скачивания веб-страниц. В отличие от браузера они не обеспечивают визуализацию информации, содержащейся в html-текстах. Программа «Spider» предназначена для загрузки веб-страниц, а программа «Crawler» осуществляет автоматический переход по всем ссылкам, найденным на странице, и определяет направление дальнейшей работы программы «Spider»;

- программа индексирования текстовых документов «Indexer» («робот-индексатор»), найденных программами «Spider» и «Crawler». Программа приписывает каждому документу совокупность ключевых слов (дескрипторов) и их весовых коэффициентов на основе статистического анализа частот этих слов в индексируемом документе и полном корпусе текстов, образованном всеми тематическими корпусами;

- база данных как хранилище поисковых образов загруженных и проиндексированных веб-документов;

- программы выдачи результатов индексирования из базы данных.

### 1.3. Подсистема аннотирования текстовых документов

Основной функцией подсистемы аннотирования является выявление в аннотируемом тексте кортежа информативных предложений и синтез аннотации. В состав подсистемы входят следующие основные структурные компоненты:

- база знаний, включающая систему словарей и совокупность тематических корпусов текстов;
- программы поиска информативных предложений. Выявляют в тексте информативные предложения и «высвечивают» их. Количество таких предложений регулируется путем задания пользователем порогового уровня информативности. При этом возможно предъявление пользователю (по его требованию) контекста каждого информативного предложения;
- программы синтеза аннотаций. Аннотация строится из информативных предложений путем поиска релевантной информации в специальной системе словарей и последующего синтеза выходного текста.

## 2. Индексирование текстовых документов

Процедура индексирования текстовых документов реализуется в два этапа. На первом этапе обрабатываются страницы html-формата: осуществляется избавление от html-тэгов и происходит структурное разбиение исходного документа. На втором этапе выявляются информативные слова и численные значения информативности для каждого слова (весовые коэффициенты). В результате получается поисковый образ документа – вектор в  $n$ -мерном евклидовом пространстве. Размерность  $n$  этого пространства на практике равна количеству словоформ в словаре системы с точностью до синонимии и словоизменения.

### 2.1. Тематические корпуса текстов

Любое непустое подмножество цепочек входного языка будем называть *текстом*, если на этом подмножестве определено отношение линейного порядка. Цепочки текста назовем *предложениями*.

Пусть имеется некоторое непустое множество текстов (совокупность текстов по конкретной тематике). Сформируем текст  $C_t$ , объединив все множества предложений каждого из этих текстов, и назовем его *тематическим корпусом* текстов. Поскольку в информационной системе представлено, как правило, несколько таких корпусов, будем обозначать их  $C_{t_i}$  ( $i$  – номер корпуса). Объединение  $C_f = \bigcup_{i=1}^n C_{t_i}$  всех тематических корпусов назовем *полным корпусом* текстов (ПКТ).

### 2.2. Вычисление информативности словоформ

Процесс выявления ключевых слов в текстовых документах включает два этапа. На первом этапе в тексте выявляются информативные словоформы, а на втором – ключевые слова (лексемы) путем корректировки частотных характеристик словоформ за счет привлечения информации из словарей синонимов и словоизменительных парадигм.

Информативность словоформы – это условная вероятность того, что данная словоформа извлечена из индексируемого текста (или релевантного ему тематического корпуса текстов)  $Td$  при условии, что она уже извлечена из ПКТ  $C_f$  [7]:

$$P(S_{Td} / S_{Cf}) = \frac{P(S_{Td} \cdot S_{Cf})}{P(S_{Cf})} = \frac{P(S_{Td}) \cdot P(S_{Cf} / S_{Td})}{P(S_{Cf})}. \quad (1)$$

В формуле (1) задействованы следующие события:

$S_{Td}$  – словоформа извлечена случайным образом из тематического корпуса текстов (или текстового документа)  $Td$  ( $Td \in C_f$ );

$S_{Cf}$  – словоформа извлечена из ПКТ  $C_f$ .

Пусть  $n_{Td}$ ,  $n_{Cf}$  – абсолютные частоты встречаемости словоформы  $a$  в индексируемом тексте (или релевантном ему тематическом корпусе текстов)  $Td$  и ПКТ  $C_f$  соответственно. Тогда

нетрудно установить [7], что при достаточно больших объемах текстов  $T_d$  и  $C_f$  формула для вычисления информативности словоформ примет вид

$$I_a = \frac{n_{Td}}{n_{Cf}}. \quad (2)$$

### 2.3. Словари базы знаний, используемые при индексировании

При индексировании текстовых документов в системе используются следующие словари базы знаний: частотный словарь словоформ, словарь синонимичных словоформ и словарь словоизменительных парадигм.

**Словарь словоформ.** Пусть  $\alpha$  – некоторая словоформа,  $P_{пкт}$  и  $P_{TKi}$  ( $i = \overline{1, n}$ ) – ее абсолютные частоты соответственно в полном и  $i$ -м тематическом корпусах текстов. Тогда совокупность кортежей типа  $\langle \alpha, P_{пкт}, P_{TK1}, P_{TK2}, \dots, P_{TKn} \rangle$  будем называть *словарем словоформ* (табл. 1).

Таблица 1

Состав и структура словаря словоформ

Словоформа	Частота в ПКТ	Частота в ТК1	...	Частота в ТК $n$	Код (номер) парадигмы
...					
бежит	0100234	0076534	...	0009987	00000021
бегут	0120410	0081123	...	0003445	00000021
...					
стол	0204055	0056534	...	0014445	00000094
стола	0401657	0074526	...	0023747	00000094
...					

В словаре словоформ каждой словоформе поставлены в соответствие:

- частота в ПКТ;
- частоты в тематических корпусах текстов (ТК1, ТК2, ..., ТК $n$ );
- номер (код) парадигмы. В первоначальном состоянии каждая словоформа словаря образует отдельную парадигму. После объединения словоформ в словоизменительные парадигмы словоформам присваивается номер парадигмы, элементом которой эта словоформа является.

**Словарь синонимичных словоформ** состоит из групп синонимичных словоформ, которые могут быть использованы при определении их информативности (табл. 2).

Таблица 2

Состав и структура словаря синонимичных словоформ

Словоформа	Синонимичные словоформы
...	
языкознание	лингвистика языковедение
...	

Словарь синонимичных словоформ создается «вручную» с использованием средств визуализации АРМ эксперта-лингвиста.

**Словарь словоизменительных парадигм** служит для поиска всех словоформ парадигмы после нахождения словоформы и ее кода в словаре словоформ. Процедура такого поиска используется при вычислении информативности слов. Словарь парадигм создается и актуализируется в человеко-машинном режиме с применением соответствующего инструментария АРМ эксперта-лингвиста. В первоначальном состоянии каждая парадигма словаря парадигм содержит одну-единственную словоформу для каждого кода словоформы. После формирования парадигм коды меняются (табл. 3 и 4).

Таблица 3

Состав и структура словаря парадигм  
(промежуточное состояние)

Код (номер) парадигмы	Словоформа
...	
00000021	бегут
00000021	бежит
...	
00000094	стол
00000094	стола
00000094	столам
...	

Таблица 4

Состав и структура словаря парадигм  
(конечное состояние)

Код (номер) парадигмы	Парадигма
...	
00000021	бегут
	бежит
...	
00000094	стол
	стола
	столам
...	

#### 2.4. Алгоритм индексирования текстовых документов

Пусть имеется некоторый текстовый документ  $Td$ . Его индексирование сводится к построению для текста  $Td$  множества пар вида  $O_{Td} = \{(a, I_a) \mid a \in W, 0 \leq I_a \leq 1\}$ , где  $W$  – множество всех различных словоформ во всех документах ПКТ  $Cf$ ;  $I_a$  – суммарная информативность словоформы  $a$ , т. е. сумма значений информативности словоформы  $a$ , всех ее словоизменений и синонимов.

Выделим из множества  $O_{Td}$  подмножество из  $l$  пар, словоформы в которых имеют наибольшую информативность (число  $l$  таких пар подбирается эмпирически), т. е. информативность всех словоформ в выделенных парах не меньше информативности словоформ во всех остальных (невыделенных) парах. Информативность словоформ во всех этих невыделенных парах положим равной нулю. Обозначим через  $I_{a_i}$  полученную информативность каждой выделенной словоформы  $a_i$ , а сформированное таким образом множество – через

$$O_{Td}^+ = \{(a_i, I_{a_i}) \mid i = \overline{1, l}\}. \quad (3)$$

Определим формально понятие поискового образа текстового документа. Введем в рассмотрение  $n$ -мерное евклидово пространство  $E$ . Для этого лексикографически упорядочим все словоформы множества  $W$ , т. е. сформируем кортеж  $W = \langle a_1, a_2, \dots, a_n \rangle$ . Для каждого индексированного текстового документа  $Td$  построим вектор в пространстве  $E$ :  $\mathbf{F}_{Td} = (I_{a_1}, I_{a_2}, \dots, I_{a_n})$ .

Вектор  $\mathbf{F}_{Td}$  будем называть *поисковым образом* текстового документа  $Td$ .

Рассмотрим алгоритм индексирования текстовых документов.

**А л г о р и т м 1.** На входе алгоритма – текстовый документ  $Td$  и количество  $l$  словоформ с ненулевой информативностью, т. е. количество ненулевых компонент вектора  $\mathbf{F}_{Td}$ , на выходе – поисковый образ документа  $\mathbf{F}_{Td}$ .

Алгоритм работает следующим образом. Выбирается очередная словоформа  $a$  документа  $Td$ , ищется в словаре словоформ, и по формуле (2) вычисляется ее информативность. Далее в словарях синонимичных словоформ и словоизменительных парадигм ищутся все синонимы и словоизменения словоформы  $a$  и для них вычисляются значения информативности. Все вычисленные значения информативности суммируются, и формируется пара  $(a, I_a)$ . После обработки всех словоформ текста  $Td$  создается его поисковый образ  $\mathbf{F}_{Td}$ . Алгоритм индексирования включает следующие шаги:

1.  $O_{Td} := \emptyset$ .
2. Выбрать очередную словоформу  $a$  из текста  $Td$  и найти ее в словаре словоформ.
3. Найти в словаре синонимичных словоформ все синонимы словоформы  $a$ .
4. Найти в словаре словоизменительных парадигм все словоизменения словоформы  $a$ .
5. Вычислить значения информативности словоформы  $a$ , всех ее синонимов и словоизменений.
6. Найти сумму  $I_a$  всех вычисленных значений информативности.
7. Поместить пару  $(a, I_a)$  в множество  $O_{Td}$ .

8. Если все словоформы текста  $Td$  исчерпаны, то перейти к п. 9, иначе – к п. 2.
9. Построить вектор  $\mathbf{F}_{Td} = (J_{a_1}, J_{a_2}, \dots, J_{a_n})$ . Конец (поисковый образ текста  $Td$  сформирован).

### 3. Поиск текстовых документов

Основными составляющими информационного поиска текстовых документов являются стратегия поиска и критерий выдачи.

#### 3.1. Стратегия поиска. Запросы. Индексирование запросов

Обозначим через  $T$  и  $Z$  некоторые непустые множества текстов, причем элементы множества  $Z$  будем называть *запросами*. Тогда один шаг поиска текстов можно промоделировать в виде частичного мультиотображения множества запросов в множество текстов:  $\pi : Z \rightarrow T$ . Частичное мультиотображение  $\pi$  назовем *поисковой функцией*.

Под стратегией поиска будем понимать последовательность его шагов, каждый последующий из которых отличается от предыдущего запросом и/или поисковой функцией. Пусть  $z_i \in Z$  – некоторые запросы, а  $\pi_i$  – поисковые функции. Тогда кортеж  $\langle \pi_1(z_1), \pi_2(z_2), \dots, \pi_l(z_l) \rangle$  будем называть *стратегией поиска*.

Стратегия поиска существенным образом зависит от типа запроса пользователя, результатов его индексирования и используемого критерия выдачи. В рассматриваемой информационной системе будем ориентироваться на два основных типа запросов:

- свободно формулируемые запросы на естественном языке. Это традиционный тип запроса. При индексировании всем его словам приписывается информативность (весовой коэффициент), равная единице. Возможно также индексирование запроса с предварительным поиском наиболее релевантного ему тематического корпуса текстов. В этом случае ключевым словам запроса ставятся в соответствие весовые коэффициенты из словаря словоформ;
- запросы, формулируемые пользователем с использованием специального графического интерфейса, где каждое слово запроса располагается на вертикальной шкале информативности (рис. 2). В зависимости от местоположения слову присваивается соответствующий весовой коэффициент.

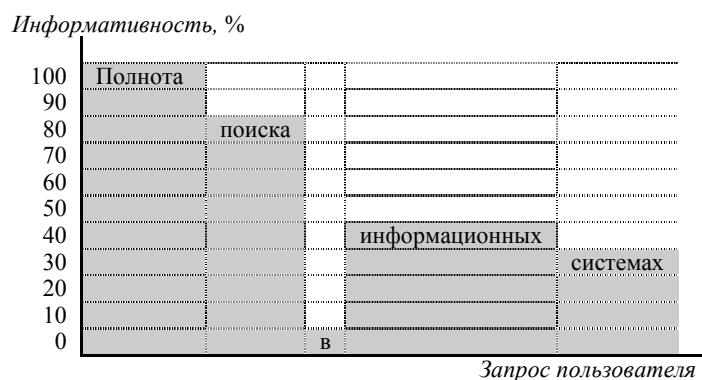


Рис. 2. Пример запроса пользователя «Полнота поиска в информационных системах» в графическом представлении

Запросы пользователя индексируются путем перевода их в векторное представление:  $\mathbf{F}_z = (J_{a_1}, J_{a_2}, \dots, J_{a_n})$ , где  $z \in Z$  – запрос,  $J_{a_i}$  – информативность слова  $a_i$ . При этом для расширения запроса используются словари синонимичных словоформ и словоизменительных парадигм. Вектор  $\mathbf{F}_z$  будем называть *поисковым предписанием*.

### 3.2. Критерий выдачи найденных документов

Под критерием выдачи понимается формальное правило, по которому вычисляется степень соответствия поискового образа документа поисковому предписанию и принимается решение о выдаче (или невыдаче) соответствующего документа пользователю. В связи с векторным представлением поисковых образов документов и поисковых предписаний в рассматриваемой информационной системе целесообразно использовать в качестве критерия выдачи косинус угла  $\varphi$  между векторами  $\mathbf{F}_{Td} = (I_{a_1}, I_{a_2}, \dots, I_{a_n})$  и  $\mathbf{F}_z = (J_{a_1}, J_{a_2}, \dots, J_{a_n})$ :

$$\cos \varphi = \frac{\mathbf{F}_{Td} \mathbf{F}_z}{|\mathbf{F}_{Td}| |\mathbf{F}_z|} = \frac{\sum_{i=1}^n I_{a_i} J_{a_i}}{\sqrt{\sum_{i=1}^n I_{a_i}^2} \sqrt{\sum_{i=1}^n J_{a_i}^2}}. \quad (4)$$

Найденные в соответствии с критерием (4) текстовые документы можно ранжировать по убыванию  $\cos \varphi$ , регулируя таким образом объем выборки.

*Примечание 1.* При программной реализации информационной системы критерий (4) целесообразно преобразовать следующим образом. Представим поисковый образ текстового документа  $Td$   $\mathbf{F}_{Td} = (I_{a_1}, I_{a_2}, \dots, I_{a_n})$  в виде выражения (3), а поисковое предписание  $\mathbf{F}_z = (J_{a_1}, J_{a_2}, \dots, J_{a_n})$  преобразуем к аналогичному виду:

$$O_z^+ = \{(a_j, I_{a_j}) \mid j = \overline{1, m}\}.$$

Сформируем множество  $O^+ = \{(b_k, I_{b_k}, J_{b_k}) \mid k = \overline{1, s}, s \leq \min(l, m)\}$ , где  $\{b_k \mid k = \overline{1, s}\} = \{a_i \mid i = \overline{1, l}\} \cap \{a_j \mid j = \overline{1, m}\}$ ;  $I_{b_k}$  – информативность слова  $b_k$  в поисковом образе документа  $Td$ ;  $J_{b_k}$  – информативность этого же слова в поисковом предписании. Другими словами,  $O^+$  – это множество ключевых слов, входящих как в поисковый образ документа, так и в поисковое предписание, с соответствующими весовыми коэффициентами. Тогда критерий выдачи (4) примет вид

$$\cos \varphi = \frac{\sum_{k=1}^s I_{b_k} J_{b_k}}{\sqrt{\sum_{i=1}^l I_{a_i}^2} \sqrt{\sum_{j=1}^m J_{a_j}^2}}.$$

### 3.3. Алгоритм поиска текстовых документов

Обозначим через  $O_T = \{\mathbf{F}_{Td} \mid Td \in T\}$  множество поисковых образов текстовых документов из множества  $T$ . Между множествами  $O_T$  и  $T$  существует взаимно однозначное соответствие. Это соответствие формализуем в виде биективного отображения  $\alpha : O_T \rightarrow T$ . Используя эти и принятые выше обозначения, рассмотрим алгоритм поиска текстовых документов.

**А л г о р и т м 2.** На входе алгоритма – множество  $T$  текстовых документов, множество  $O_T$  их поисковых образов, биективное отображение  $\alpha$ , запрос пользователя  $z \in Z$  и поисковая функция  $\pi : Z \rightarrow T$ , на выходе – кортеж найденных и ранжированных текстовых документов  $\pi(z)$ .

1.  $M := \emptyset, K := \emptyset$ .

2. Индексировать запрос пользователя  $z \in Z$ . В результате индексирования запроса получаем поисковое предписание  $\mathbf{F}_z = (J_{a_1}, J_{a_2}, \dots, J_{a_n})$ .

3. Найти в множестве  $O_T$  поисковый образ документа  $\mathbf{F}_{Td} = (I_{a_1}, I_{a_2}, \dots, I_{a_n})$ , для которого  $\cos \varphi \neq 0$  (формула (4)). Поместить поисковый образ  $\mathbf{F}_{Td}$  в множество  $M$ .

4. Если все поисковые образы документов, для которых  $\cos \varphi \neq 0$ , в множестве  $O_T$  найдены, то перейти к п. 5, иначе – к п. 3.

5. Упорядочить все поисковые образы документов множества  $M$  по убыванию значений  $\cos \varphi$  (формула (4)) и поместить их в кортеж  $K = \langle \mathbf{F}_{Td_1}, \mathbf{F}_{Td_2}, \dots, \mathbf{F}_{Td_r} \rangle$ .

6. Выдать найденные документы  $\langle \alpha(\mathbf{F}_{Td_1}), \alpha(\mathbf{F}_{Td_2}), \dots, \alpha(\mathbf{F}_{Td_r}) \rangle$ . Конец.

*Примечание 2.* Если при индексировании запросов пользователя информационной системы применять тематический корпус текстов, сформированный пользователем, то таким образом можно реализовать поиск с адаптацией к его информационным потребностям.

#### 4. Аннотирование найденных текстовых документов

Процесс аннотирования текста включает два основных этапа. На первом этапе в аннотируемом тексте выявляется совокупность информативных предложений. На втором этапе с использованием специальных словарей базы знаний синтезируется аннотация путем конкатенации (объединения) фразеологических словосочетаний и ситуативных связей между ними.

##### 4.1. Словари базы знаний, используемые при аннотировании

Для реализации процесса аннотирования текстовых документов в системе предусмотрены следующие словари базы знаний: словарь прагматически полных синтагматических структур (ПП-структур), словарь синонимичных ПП-структур, право- и левосторонний словарь ситуативных связей, а также ситуативно-синтагматический словарь.

*Словарь ПП-структур.* ПП-структура – это информативная в некотором тематическом корпусе текстов фраза (как правило, именное словосочетание), выражающая законченную по смыслу формулировку понятия. Словарь ПП-структур (табл. 5) создается в полуавтоматическом режиме с использованием программно-сформированного исходного файла и средств визуализации АРМ эксперта-лингвиста.

Таблица 5  
Состав и структура словаря ПП-структур

ПП-структура
...
гидрофизический институт
...
дистанционное зондирование Земли
...

*Словарь синонимичных ПП-структур.* Структура словаря (табл. 6) аналогична структуре словаря синонимичных словоформ (см. табл. 2). Она создается «вручную» с использованием средств АРМ эксперта-лингвиста.

Таблица 6  
Состав и структура словаря синонимичных ПП-структур

ПП-структура	Синонимичные ПП-структуры
...	...
дистанционное зондирование Земли	аэрокосмическая съемка Земли дистанционное исследование Земли
...	...

*Правосторонний словарь ситуативных связей.* Под ситуативными связями будем понимать подцепочки предложений языка, предшествующие ПП-структуре или следующие за ней, а также связывающие ПП-структуры между собой. Обозначим через *Sit* множество ситуативных связей, а через *Str* – множество ПП-структур языка. Тогда отношение ЛП, определенное на паре множеств *Sit*, *Str*, будем называть отношением *правосторонней конкатенации*. На практике отношение ЛП реализуется в виде правостороннего словаря ситуативных связей (табл. 7).

Таблица 7  
Состав и структура правостороннего словаря ситуативных связей

ПП-структура и ее родовые понятия	Правосторонняя ситуативная связь	Наличие связи в левостороннем словаре
...	...	...
Космическая фотоаппаратура ↑Такая аппаратура↑ ↑Аппаратура↑	используется для является важной составляющей изготовлена на	+ – +
...	...	...
Суп из ласточкиных гнезд	является любимым лакомством остался невостребованным.	– –
...	...	...



В первой позиции каждой записи правостороннего словаря ситуативных связей содержится ПП-структура, а во второй – ситуативные связи, каждая из которых может быть присоединена справа к ПП-структуре. В третьей позиции находится признак наличия данной ситуативной связи в левостороннем словаре. Родовые понятия, заключенные между символами «↑...↑», при синтезе выходных предложений вставляются вместо соответствующих ПП-структур. Обычно это происходит в случаях, когда одна и та же ПП-структура содержится в соседних предложениях.

*Левосторонний словарь ситуативных связей.* Отношение ПА, определенное на паре множеств  $Str, Sit$ , будем называть отношением *левосторонней конкатенации*. В подсистеме автоматического аннотирования отношение ПА представлено левосторонним словарем ситуативных связей (табл. 8). Каждая запись этого словаря включает две позиции. В первой позиции представлена ситуативная связь, а во второй – ПП-структуры, каждая из которых может быть присоединена в соответствующем контексте к ситуативной связи из первой позиции словаря.

Таблица 8

Состав и структура левостороннего словаря ситуативных связей

Левосторонняя ситуативная связь	ПП-структура и ее родовые понятия
...	...
используется для	дистанционного зондирования Земли. ↑зондирования.↑такого исследования.↑
	аэрокосмической съемки местности ↑подобной съемки.↑
	синтаксического анализа текста.
	приготовления супа из ласточкиных гнезд.
...	...
Предлагается новый подход к решению проблемы	четырёх красок.
	дистанционного исследования Земли.
	принятия решений в условиях неопределенности.
...	...

*Ситуативно-синтагматический словарь.* Пусть по-прежнему  $Ct_i$  ( $i = \overline{1, n}$ ;  $n \geq 2$ ) – тематические корпуса текстов,  $Cf = \bigcup_{i=1}^n Ct_i$  – ПКТ, а  $Sint$  – множество всех синтагматических структур ПКТ  $Cf$ . Введем в рассмотрение для каждого корпуса текстов  $Ct_i$  ситуативное отношение.

Отношение толерантности  $\Psi_i$  (рефлексивное и симметричное бинарное отношение) на множестве  $Sint$  назовем *ситуативным отношением* в корпусе текстов  $Ct_i$ , если любая упорядоченная пара синтагматических структур  $(\mu, \nu)$  из множества  $Sint$  является элементом отношения  $\Psi_i$  тогда и только тогда, когда вероятность совместной встречаемости структур  $\mu$  и  $\nu$  в тематическом корпусе текстов  $Ct_i$  не меньше некоторого порогового значения (*уровня* ситуативной связи). Объединение  $\Psi = \bigcup_{i=1}^n \Psi_i$  будем называть ситуативным отношением в ПКТ  $Cf$ . Под совместной встречаемостью двух синтагматических структур здесь понимается наличие этих структур (или их синонимов) в одном и том же предложении тематического корпуса  $Ct_i$ . При этом словоформы синтагматических структур отождествляются с точностью до словоизменения, зафиксированного в словаре словоизменительных парадигм.

В информационной системе ситуативное отношение  $\Psi$  в ПКТ целесообразно представить в виде ситуативно-синтагматического словаря (табл. 9).

Таблица 9

Состав и структура ситуативно-синтагматического словаря

ПП-структура	ПП-структура	Частота в $Cf$
...	...	...
аэрокосмические съемки	околоземная фотоаппаратура	0223651
...	...	...
дистанционное зондирование Земли	космические аппараты	0034998
...	...	...
жители Тайланда	суп из ласточкиных гнезд	0005643
...	...	...

#### 4.2. Алгоритм синтеза аннотаций

В основу алгоритма автоматического аннотирования текстовых документов положены следующие основные принципы:

*информативности аннотации* – аннотация текстового документа строится на основе совокупности его информативных предложений (в данном случае всех предложений, содержащих информативные слова);

*политематичности аннотации* – аннотация должна отражать в общем случае все тематические аспекты аннотируемого текста, которые должны быть представлены в нем как совокупность тематических разделов. При выделении этих разделов используется ситуативно-синтагматический словарь;

*мономатичности тематического раздела аннотации* – каждый тематический раздел аннотации должен соответствовать тематическому аспекту аннотируемого текста;

*содержательной достаточности аннотации* – аннотация должна иметь объем, не больший того, который обеспечивает уяснение тематики аннотируемого документа;

*связности аннотации* – реализуется путем использования словаря ПП-структур, а также право- и левостороннего словарей ситуативных связей.

Рассмотрим алгоритм синтеза аннотаций.

**А л г о р и т м 3.** На входе алгоритма – текстовый документ  $Td$ , на выходе – аннотация.

1. Выявить в тексте  $Td$  информативные слова по формуле (1). (Обозначим через  $T_1$  множество всех предложений, содержащих информативные слова.)

2. С использованием словаря ПП-структур и выделенных в тексте  $T_1$  информативных слов найти в  $T_1$  ПП-структуры для всех выделенных слов.

3. Исключить из текста  $T_1$  все предложения, в которых не найдена ПП-структура хотя бы для одного выделенного слова. (Обозначим полученный текст через  $T_2$ .)

4. Поместить все предложения из текста  $T_2$  в файл  $FT_2$  (табл. 10). В последний столбец табл. 10 занести частоту встречаемости пар ПП-структур из ситуативно-синтагматического словаря.

Таблица 10

Фрагмент файла  $FT_2$  частично структурированных предложений

Цепочка из $T_2$	ПП-структура	Цепочка из $T_2$	ПП-структура	Цепочка из $T_2$	Частота в $Cf$
...					
–	–	Речь идет о новом подходе к решению проблемы	дистанционного зондирования Земли	упомянутыми выше средствами.	–
Реализация процесса	дистанционного зондирования Земли	основана на использовании средств	околоземной фотоаппаратуры	, расположенной на орбите.	0233450
...					

5. Разбить файл  $FT_2$  частично структурированных предложений на тематические группы следующим образом. Выбрать произвольную запись (пару ПП-структур) из файла  $FT_2$ . Затем выбрать все пары ПП-структур, одна из которых содержится в первой паре. Далее процедуру повторить для каждой из последних выбранных пар. Процесс заканчивается в одном из двух случаев: 1) когда будут выбраны все пары ПП-структур из данного файла; 2) когда не будет найдена ни одна пара ПП-структур, удовлетворяющая указанному выше условию (этот факт является признаком выделения тематической группы). Другие тематические группы выделяются аналогичным образом. На выходе шага 5 – файл  $FT_3$  (табл. 11). Порядок следования предложений в  $FT_3$  соответствует их порядку в файле  $FT_2$ .

Таблица 11

Фрагмент файла  $FT_3$  частично структурированных предложений с выделенными тематическими группами

Цепочка из $T_2$	ПП-структура	Цепочка из $T_2$	ПП-структура	Цепочка из $T_2$	Частота в $Cf$
<i>Тематическая группа 1</i>					
...					
Реализация процесса	дистанционного зондирования Земли	основана на использовании средств	околоземной фотоаппаратуры	, расположенной на орбите.	0233450
...					
<i>Тематическая группа 2</i>					
...					
Рассмотрены	геоинформационные системы	, которые предназначены для обработки	космических снимков	, переданных с искусственного спутника Земли.	0005643
...					

6. Синтезировать группы предложений следующим образом. Выбрать очередное предложение из файла  $FT_3$ , например предложение из тематической группы 1 (табл. 11). Первая пара фраз «Реализация процесса; дистанционного зондирования Земли» ищется в левостороннем словаре ситуативных связей (см. табл. 8). При этом ситуативная связь, релевантная фразе «Реализация процесса», должна быть началом предложения (т. е. начинаться с прописной буквы). В результате получим следующую ситуативную связь и соответствующую ей ПП-структуру:

При реализации	дистанционного зондирования Земли ↑зондирования↑
----------------	---

Далее в правостороннем словаре ситуативных связей (см. табл. 7) следует искать пару фраз «дистанционного зондирования Земли; основана на использовании средств». В результате имеем релевантную запись правостороннего словаря ситуативных связей

дистанционного зондирования Земли ↑зондирования↑	используются средства
---	-----------------------

На завершающем этапе обработки предложения в левостороннем словаре ситуативных связей найти фрагмент «используются средства; околоземной фотоаппаратуры». Найденную информацию поместить в файл  $FT_4$  (табл. 12). Процесс продолжать для всех предложений файла  $FT_3$ . Аналогичную процедуру выполнить для всех тематических групп.

Таблица 12

Фрагмент файла  $FT_4$  групп предложений

Ситуативная связь	ПП-структура	Ситуативная связь	ПП-структура	Ситуативная связь	Частота в $Cf$
<i>Тематическая группа 1</i>					
...					
При реализации	дистанционного зондирования Земли ↑зондирования↑	используются средства	околоземной фотоаппаратуры.	—	0233450
...					
<i>Тематическая группа 2</i>					
...					
—	Геоинформационные системы ↑Эти системы↑	предназначены для обработки	космических снимков.	—	0005643
...					

7. Синтезировать аннотацию следующим образом. Вначале определить число тематических групп (например, две группы). Если групп две или больше, то выдать сообщения вида: «Работу можно условно разделить на две части. В первой части работы представлена следующая информация». Далее приводятся предложения из первой тематической группы: «Предлагается новый подход к решению проблемы дистанционного зондирования Земли. При реализации зондирования используются средства околоземной фотоаппаратуры. Этой аппаратурой

снабжены космические аппараты...». Далее выдать сообщение: «Во второй части работы представлена следующая информация». Далее следуют предложения из второй группы: «Геоинформационные системы предназначены для обработки космических снимков. При обработке используются специальные алгоритмы распознавания...». Конец.

### Заключение

Предложенные в статье алгоритмы могут быть использованы для индексирования, поиска и аннотирования научно-технической информации как в Интернете, так и в локальных базах данных. Благодаря использованию ситуативно-синтагматического словаря возможен синтез аннотаций текстовых документов на различных выходных языках при наличии в информационной системе двуязычных словарей ПП-структур и ситуативных связей. При соответствующем подборе тематики и иерархической структуры корпусов текстов возможны поиск и аннотирование текстовых документов с адаптацией к информационным потребностям пользователя и с учетом стилистической окраски документов (например, публицистика, художественная или научно-популярная литература).

### Список литературы

1. Новые требования к издательской аннотации / Азбука библиографа [Электронный ресурс]. – 2004. – № 12 – Режим доступа : [http://www.book.ru/?page=10&mode=c&origin=home&id\\_a=314](http://www.book.ru/?page=10&mode=c&origin=home&id_a=314). – Дата доступа : 23.12.2008.
2. Воройский, Ф.С. Информатика. Новый систематизированный толковый словарь-справочник (Введение в современные информационные и телекоммуникационные технологии в терминах и фактах) / Ф.С. Воройский. – М. : Физматлит, 2003. – 760 с.
3. Тактаев, С. Поиск информации в компьютерных сетях: новые подходы / С. Тактаев [Электронный ресурс]. – Режим доступа : <http://www.searchengines.ru/articles/004603.html>. – Дата доступа : 10.02.2009.
4. Храмцов, П. Информационно-поисковые системы Internet / П. Храмцов [Электронный ресурс]. – Режим доступа : <http://www.osp.ru/os/1996/03/178885/>. – Дата доступа : 10.02.2009.
5. Бэлэни, Н. Будущее Web – за семантикой / Н. Бэлэни [Электронный ресурс]. – Режим доступа : <http://www.iso.ru/journal/articles/432.html>. – Дата доступа : 10.02.2009.
6. Соколова, Е.Г. Автоматическая генерация текстов на ЕЯ (портрет направления) / Е.Г. Соколова, М.В. Болдасов [Электронный ресурс]. – Режим доступа : <http://www.dialog-21.ru/Archive/2004/Sokolova.htm>. – Дата доступа : 24.06.2008.
7. Липницкий, С.Ф. Семантический анализ текста на основе ситуативно-синтагматической сети / С.Ф. Липницкий // Информатика. – 2005. – № 2. – С. 102–110.

Поступила 10.02.09

*Объединенный институт проблем  
информатики НАН Беларуси,  
Минск, Сурганова, 6  
e-mail: lipn@newman.bas-net.by*

**S.F. Lipnitsky, A.A. Mamchich, S.A. Sorudeykina**

### WEB-SEARCH AND ANNOTATION OF SCIENTIFIC AND TECHNICAL INFORMATION ON THE BASIS OF THEMATIC CORPUSES

Algorithms for searching textual documents in the Internet and their annotations are suggested. The architecture of a system for web-search and annotation of the scientific and technical information is designed. Composition and structure of the linguistic dictionaries knowledgebase used for indexing text documents and annotation synthesis are analyzed.